ED 358 114 TM 019 880

AUTHOR Wise, Lauress

TITLE Scoring Rubrics for Performance Tests: Lessons Learned from Job Performance Assessment in the Military.

INSTITUTION Defense Manpower Data Center, Monterey, CA.

PUB DATE Apr 93

NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Educational Assessment; Educational Research; Evaluation Methods; Generalizability Theory; Industrial Psychology; *Job Performance; *Military Personnel; *Occupational Tests; Organizational Development; *Performance Tests; Personnel Evaluation; Personnel Selection; *Scoring; Standards; Test Construction; Training

IDENTIFIERS Department of Defense; Job Performance Measurement Project; *Performance Based Evaluation

ABSTRACT

Industrial and organizational psychologists for the Department of Defense have been working for the past 10 years to develop high fidelity measures of job performance for use in validating job selection procedures and standards. Information on developing and scoring performance exercises in the Job Performance Measurement (JPM) Project is presented, and lessons that might be useful in education are extracted. In many ways, the task of the industrial psychologist is easier than that of the educator because of broader agreement about how the task should be performed and close alignment between training and expected performance. Tasks identified by each Armed Service were analyzed, and scoring rules were developed. The following lessons seem especially pertinent to educational assessment: (1) careful specification of the domains assessed is essential for evaluating the adequacy of any sample selected; (2) scoring elements that assess adherence to processes that are taught will have better diagnostic value (and possibly greater validity) than will those that just reflect the quality of output; (3) scoring procedures must be anchored to observable criteria; and (4) generalizability theory provides a useful framework for evaluating alternative scoring rubrics. One table lists the JPM occupational specialties, and two figures illustrate the discussion. An attachment summarizes the lessons to be learned. (SLD)

Scoring Rubrics for Performance Tests: Lessons Learned from Job Performance Assessment in the Military

Dr. Lauress Wise
Defense Manpower Data Center

Paper presented at the Annual Convention of the National Council for Measurement in Education, April 14, 1993, Atlanta, GA.

NOTE: The views expressed are those of the author and do not necessarily reflect the position of the Defense Manpower Data Center or other agencies within the Department of Defense.

DEFENSE MANPOWER DATA CENTER
99 Pacific Street, Suite 155A • Monterey, CA 93940

Scoring Rubrics for Performance Tests:
Lessons Learned from Job Performance Assessment in the Military

Over the past several years, there has emerged an increasingly universal consensus that: (1) many of the skills we want our schools to impart to young people are not well measured by traditional multiple-choice tests, and (2) universal reliance on multiple choice tests encourages teaching and study habits that are at odds with deeper educational goals. As a consequence, educational assessments at the national, state, and local level are incorporating a wide variety of "performance-based" assessment strategies designed more to assess proficiency at practiced skills than to measure simpler factual knowledge.

Performance-based assessment is a somewhat new endeavor in educational research, but it has been studied for some time in other arenas. Measurement of job performance has long been an area of concern for industrial psychologists where strategies such as "work sample" assessments are not new. For the past ten years, industrial and organizational psychologists working for the Department of Defense have been engaged in an unprecedented effort to develop high fidelity measures of job performance for use in validating job selection procedures and standards (Wigdor & Green, 1991). This effort, known as the Job Performance Measurement (JPM) Project, provides a number of lessons that may be useful in the development of performance-based educational assessments. The general goal of this paper is to present information on the approach to developing and scoring performance exercises used in the JPM Project and to suggest lessons that might also be useful in an educational arena.

Background

The JPM Project was a coordinated effort involving each of the Armed Services, overseen by the Office of the Assistant Secretary of Defense for Force Management and Personnel (OASD-FM&P). Each Service selected a sample of jobs to represent the broad domain of occupational specialties available to their recruits. Although the Services organize their entry jobs somewhat differently, four broad categories may be used to categorize jobs in terms of their cognitive requirements: (1) Mechanical, (2) Administrative, (3) Electrical/Electronic, and (4) General. Job classification systems used by the Air Force and Marine Corps use these four categories. The Army further divides the General category into combat, technical, communications, operators, and general maintenance. The Navy also has further subdivisions within the Mechanical and Electrical families. The sample of jobs selected by the Services represented the different job categories used in each of their selection and classification systems. Table 1 lists the jobs included in the JPM project by Service and by job category.

Hands-On Performance Tests (HOPT). Extensive job analyses were performed for each of the jobs included in the JPM study, resulting in lists of roughly 500 to 1,000 "tasks" identified in training materials, job-specific manuals, or occupational surveys. Tasks in the initial task lists varied considerably in detail and complexity. Simpler tasks were combined and more peripheral tasks were eliminated to narrow the initial list down to a revised list of from 150 to 250 tasks. Subject-matter-experts (SMEs) rated the criticality, difficulty, and frequency of each of the tasks and sorted them into separate content groups. A final sample of tasks was selected from the revised list for each job based on the information provided by the SMEs. The final number of tasks varied somewhat by Service. The Army selected 15 tasks for each job. The Marine Corps project included as many as 35 distinct tasks for one job. The final sample for each job was designed to emphasize the most critical and frequently performed tasks, to cover each of the different content areas identified in the task sorting exercises, and to avoid tasks that were so easy or difficult that little information on individual differences could be gained. For some jobs, an assessment of testing feasibility also was used in deciding among equally relevant tasks.

Once a set of tasks had been selected, scoring rules were developed for each of the target tasks. Critical steps or behaviors were identified in consultation with SMEs. For the most part, each step or behavior was scored as "GO" if it was performed successfully and "NO-GO" if it was not. In some cases, more quantitative information (e.g., typing speed, time for completion, number of targets hit) was also incorporated. In most cases, the total score for the task was the percentage of steps/behaviors performed successfully. Figure 1 shows an example of a scoring sheet for one of the truck-driver tasks.

In addition to developing the specific scoring rubric for each task, the project team developed a program for training scorers on the use of the scoring rubric. Generally, tasks were organized into four to eight stations, with a different scorer assigned to each station. Scorers received up to two days of training in scoring the tasks for the station(s) to which they were assigned. Special studies were also conducted that included use of multiple scorers at each station and planned rotation of scorers through all stations.

Rating Scales. An independent approach to job analyses was used in developing the performance-rating scales. The critical incident technique (Flanagan, 1954) was used to elicit examples of specific behaviors judged by SMEs to be particularly effective or ineffective. Hundreds of incidents were collected for each job, edited into a common format, and clustered into 8 to 15 categories. SMEs then participated in "retranslation" exercises where the edited incidents were matched to preliminary category descriptions and rated with regard to the level of effectiveness that they exemplified. Categories were combined or eliminated where there was some ambiguity about the matches of incidents to categories.

Incidents for which there was high agreement as to the category and level of performance they exemplified were considered for use as scale anchors. A separate rating scale was constructed for each of the final incident categories. For each scale, one or more incidents were summarized to describe particularly effective performance (scale levels 6 and 7), particularly ineffective performance (scale levels 1 and 2), and average performance (levels 3, 4, and 5). In some cases there were not many incidents at the middle levels of performance, so the behavioral summaries for the middle performance level were synthesized to be consistent with the summaries at either end of the scale. Figure 2 provides an example of one of the rating scales that resulted from the process.

In addition to developing well-anchored behavioral scales, the project staff also developed a rater-training program (Pulakos & Borman, 1986). This training program was designed to reduce halo effects and to promote more accurate comparisons among those rated, building on the results of previous research on similar programs (Borman, 1979).

Wise (1992) suggested some lessons for the design of educational assessments that might be drawn from the JPM endeavor. (See Attachment A.) The remainder of this paper focuses on the issue of how assessments are scored.

Lessons Learned

The lessons learned from job performance measurement about scoring rubrics may be organized under three general questions:

1. How were the assessment exercises designed or selected?

2. What different types of scoring rubrics were used with the exercises that were selected?

3. How was the adequacy of these rubrics evaluated?

"Lessons learned" relevant to each of these three questions are described in the remainder of this paper.

Development of Performance Tasks

The really important question for performance assessments is not so much *how* to score as *what* to score. If inappropriate exercises are administered, the best scoring rubrics in the world will not yield a valid assessment of the target ability. The steps used in the JPM project to identify sets of performance tasks provide an example of how meaningful performance tasks might be derived, although some elements of this approach may not translate easily into the educational assessment environment. A brief description of the steps used in developing the hands-on tests is provided here.

Careful specification of the domain to be assessed. The development of performance tests for military jobs was relatively simple because results from extensive job analyses were available and because of a close alignment of job training and job performance. A "manual" was available for each job listing all of the tasks that had to be performed. Training material also was available for each of the tasks in these manuals. Thus, the domain of "job performance" had already been divided into a discrete set of behaviors or tasks, and it was possible to possible to create an essentially exhaustive list of these elements that constituted job performance.

Involvement of subject-matter-experts (SMEs). After some editing to even out the level of detail encompassed by different tasks, surveys were conducted to assess the frequency and criticality of each task and to develop a grouping of tasks into related clusters. Both trainers and supervisors who had extensive knowledge of the domain of interest were available. For each job, there was a school designated as the "proponent" for that job, which provided input to the task definition process and also signed off on the final results.

Systematic sampling from the target domain. Sampling was then conducted within each cluster to ensure coverage of all of the different "types" of tasks with priority given to more frequent and more important tasks. In the educational arena, curricular frameworks serve somewhat the same function of specifying the domain to be covered in an assessment, but there is no direct counterpart of the task lists that provided an enumeration of elements in the domain of interest.

Application to educational assessment. The curricular frameworks developed for educational assessments serve the general purpose of specifying the domain of performance to be assessed. Continuing efforts are required, however, to achieve better agreement on how to characterize all of the knowledge, processes, and behaviors that constitute the specific elements of these assessment domains. While "SMEs" are plentiful in the educational arena, there is, unfortunately, no central authority that has the definitive word on the content of each curricular area. Professional organizations, such as the National Council of Teachers of Mathmatics (NCTM), have made progress in developing curricular frameworks for use with educational assessments, but the process is cumbersome and universal consensus is nearly impossible. Without a reasonably comprehensive specification of the assessment domains, it is very difficult to determine the extent to which scores derived for specific exercises generalize to the larger domain of the assessment.

Development of Scoring Procedures

Two different kinds of scoring procedures were developed in the JPM Project. The scoring procedures for the hands-on performance tests would appear to be most directly relevant to the development of scoring rubrics for educational performance assessments, but the behaviorally-anchored rating scales may actually be more similar to the type of scoring rubrics most commonly used in such assessments. In both cases, several questions were addressed in developing scoring procedures. Each of these questions and their impact on scoring procedures are described briefly here.

Is there a right answer? The first step in developing scoring guidelines for the hands-on tests was to determine the "correct" procedure for performing each task. In the military performance arena, this was a relatively simple endeavor as there were manuals and training materials that provided definitive descriptions of correct procedures. In the educational arena, life is not nearly as simple. Indeed, the move to authentic performance assessments reflects, in part, a desire to move away from an overly simplistic, binary view of the world where all responses are either correct or incorrect. Consider the following statement from a description of California's science assessment:

> "In a performance assessment students are encouraged to demonstrate understanding by conducting an investigation, collecting and analyzing data, and forming a conclusion. These types of assessments have no prescribed answer, but allow for a variety of appropriate student responses, including writing, drawing, and/or manipulation of data. In order to accommodate a wide range of responses, as well as to encourage the evaluating of the students entire thinking process, holistic scoring guides or rubrics were developed for all tasks" (California State Department of Education, 1993).

Another way of considering the issue of one correct, versus many nice, answers is to ask whether we are assessing performance of a procedure that is specifically taught. In the job performance arena, students were taught to follow and encouraged to practice a specific procedure in performing each task. Part of the paradigm shift in educational assessment has been a desire to develop tests that teachers could and should teach to, and to encompass tasks that students might specifically practice. In reality, however, the new educational assessments tend more toward novel tasks that require students to generalize from specific knowledge and procedures that they have been taught. Consider a common beginning to writing assessment prompts: "Compare and contrast ... ." Did anyone ever teach you a specific procedure to be followed in responding to such a task? How about procedures for responding to a prompt such as "Describe your favorite music."

Where there are many correct ways of responding to an assessment exercise, the performance-rating scales be more relevant for educational assessments than the hands-on tests. The critical incident technique used in developing anchors for these scales might be used in developing examples of particularly effective or ineffective responses. Perhaps scoring procedures similar to those used with diving or ice skating competitions might be developed where guidelines indicate how many points are to be subtracted for various kinds of defects in performance. The hands-on scoring procedures might be viewed from this perspective. The "NO-GO" marks for specific steps were a form of "points off" for bad behaviors.

Are we more concerned with process or output? A second question addressed in developing scoring procedures is whether we want to measure adherence to the procedure that is followed or to judge the quality of the output that is produced. The hands-on performance measures from the JPM Project included both types of criteria. In the example shown in Figure 1, some of the scoring elements, such as "keeping both hands on the wheel" reflected adherence to procedures that had been taught, while other steps, such as "shifted gears without grinding" reflected outcome more than process. In many cases where there is a clearly correct procedure, judging the process and judging the outcome are equivalent. In the educational arena where there are not always correct procedures, we are forced to rely more on judgment of outcome or product. This may be unfortunate, as it makes it more difficult to link assessment results to instruction designed to improve student performance.

Is adherence to prescribed procedures observable? In order for scoring to be reliable, it is important that what is scored be readily observable by the scorer. One of the primary reasons for scoring output rather than process was that output was always observable, while the process used to develop the output may not have been. "Shifting gears without grinding", for example, requires appropriate changes in tension on the clutch as the shift proceeds. The outcome of grinding gears if very easy to observe (hear), while the process used to produce or avoid this grinding is not.

Application to educational assessment. In educational assessment, we are often backed into rubrics for "holistic" scoring of examinee "output" because: (1) there may not be agreement on the correct process for producing the output, (2) it is too difficult to make adherence to the process observable, and/or (3) it is simply too costly to make many detailed scoring judgments about each exercise. Nonetheless, there are plenty of examples, such as the "show your work" and "partial credit" approach for mathematics problems, where more detailed scoring of adherence to process can be used. In the JPM arena, the hands-on scoring procedures were generally considered the ideal because they captured most precisely whether the students were following what they were taught and they provided good diagnostic information. It also seems reasonable to propose that a detailed scoring of adherence to process be the ideal for educational assessments as well, particularly where the assessment and instruction are intertwined.

Where it is necessary to fall back on holistic scoring of output, the process used in developing performance rating scales may be useful to consider. This process involves expert "focus" groups given specific prompts to elicit critical aspects of effective and ineffective performance. Similar efforts might be employed to identify differentiating characteristics of good and bad responses to educational assessment exercises.

Evaluation of scoring procedures.

A final question to consider about scoring rubrics is "How do we know a good rubric when we see one?" There has been considerable debate in the educational community about the extent to which traditional psychometric criteria should be used in evaluating performance-based assessments. Some argue that the impact of the assessment on teaching practices and study habits is more important than the information conveyed in specific scores derived from the assessment. One thesis of this paper is that a more detailed scoring of adherence to instructed procedures is generally preferable to a holistic scoring of output. How can we determine whether this is the case?

Much of the debate over psychometric issues reflects an unfortunate concern over the primacy of reliability or validity. Proponents associated with traditional psychometrics argue that assessment results can't be valid if they are not reliable, while proponents of new forms of assessment argue that it does not matter whether it is reliable if it is not what we really want to measure (valid). The difference is that reliability is primarily a statistical issue and can be determined analytically, while validity begins with judgment about what is to be measured.

The appropriate criteria for evaluating scoring rubrics will, of course, depend to a large extent on how the resulting scores are to be used. If the goal of the assessment if primarily instructional, the psychometric characteristics of the scores may not be as important as the impact of the assessment on student attitudes, beliefs, and practices. If, however, the scores are to be used diagnostically or in a "high stakes" evaluation of the student, the teacher, or the school system, then both the validity and the reliability of the scores are critical.

When psychometric considerations are important, generalizability theory provides a comprehensive framework for assessing reliability and, to a certain extent, validity (Webb et al., 1989; Shavelson et al., 1990). In contrast to a single reliability coefficient, the generalizability approach tells us the degree to which scores are consistent across scorers and occasions, and across different exercises.

In the absence of more ultimate criteria, validity must be judged in terms of the specification of the domain from which the exercises are drawn and the degree to which scores for a sample of exercises will generalize to the whole domain. Specification of the domain of performance to be assessed was where we started this discourse.

Summary

This paper has been an exercise in comparing and contrasting the assessment of job performance as developed by industrial psychologists and new trends in performance-based educational assessments. In many ways, the task of the industrial psychologists was much easier. The domain of behaviors to be assessed was well specified, there was general agreement on how each task should be performed, and there was a close alignment between how examinees were trained and how they were expected to perform on the tests. In addition, one-on-one testing was feasible for the samples used in the JPM study. Nonetheless, several lessons from these efforts are worth consideration in the development of scoring procedures for educational assessments, including the following:

1. Careful specification of the domain of behaviors being assessed is essential to evaluating the adequacy of any particular sample selected for use in an assessment.

2. Scoring elements that assess adherence to processes that are taught rather than just the quality of the output from these processes will have better diagnostic value and, perhaps, greater validity.

3. Scoring procedures must be anchored to observable criteria, so efforts to make adherence to prescribed practices observable are useful.

4. Generalizability theory provides a useful framework for evaluating alternative scoring rubrics.

References

Borman, W.C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 412-421.

California State Department of Education. (1993). Science: New Dimensions in Assessment. Sacramento, CA: California State Department of Education.

Campbell, C.H., Campbell, R.C., Rumsey, M.G., & Edwards, D.C. (1986). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, C.H., Ford, P., Rumsey, M.G., Pulakos, E.D., Borman, W.C., Felker, D. B., de Vera, M. V., & Reigelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. Personnel Psychology, 43, 277-300.

Campbell, J.P. (Ed.). (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J.P., McHenry, J.J., & Wise L. L. (1990). Modeling job performance in a population of jobs. Personnel Psychology, 43, 313-334.

Flanagan, J.C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Gottfredson, L. (1990). The evaluation of alternative measures of job performance. In A.K. Wigdor & B.F. Green (Eds.), Performance assessment for the workplace, Vol. 2. Washington, DC: National Academy Press.

Harris, D.A., McCloy, R.A., Dempsey, J.R., Roth, C., Sackett, P.R., Hedges, L.V., Smith, D.A., & Hogan, P.F. (1991). Linking enlistment standards to job performance, Phase I: A job performance model. Alexandria, VA: Human Resources Research Organization.

McHenry, J.J., Hough, L.M., Toquan, J.L, Hanson, M.A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.

Pulakos, E.D., & Borman, W.C. (1986). Development and field test of Army-wide rating scales and rater orientation and training program (ARI Technical Report 716). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shavelson, R.J., Mayberry, P.W., Weichang, L., & Webb, N.M. (1990). Generalizability of job performance measurements: Marine Corps rifleman. Military Psychology, 2, 129-144.

Web, N.M., Shavelson, R.J., Kim, K.S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. Military Psychology, 1, 91-110.

Whetzel, D. L. (1991). Multidimensional screening: Comparison of a single-stage personnel selection/classification process with alternative strategies. Unpublished Doctoral Dissertation, George Washington University, Washington DC.

Wigdor, A. K., & Green, B. F. (Eds.) (1991). Performance Assessment for the Workplace. Washington, DC: National Academy Press.

Wise, L.L., McHenry, J.J., & Campbell, J.P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. Personnel Psychology, 43, 355-366.

Wise, L.L. (1992). Lessons learned from military performance assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wise, L.L, Peterson, N.G., Hoffman, R.G., Campbell, J.P., & Arabian, J.M. (1991). Army Synthetic Validity Project: Report of Phase III results. (ARI Technical Report 922). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Table 1. Occupational Specialties Included in the Job Performance Measurement Project

Service	Job Family	Occupational Title
Army	General	Infantryman
	General	Cannon Crewman
	General	Tank Crewman
	General	Radio Teletype Operator
	Administrative	Medical Specialist
	Mechanical	Light Wheel Vehicle/Power Generator Mechanic
	General	Motor Transport Operator
	Administrative	Administrative Specialist
	General	Military Police
Navy	Mechanical	Machinists Mate
	General	Radioman
	Electrical	Electronics Technician
	Electrical	Electrician's Mate
	General	Fire Control Specialist
	Mechanical	Gas Turbine Technician, Mechanical
Air Force	Mechanical	Jet Engine Mechanic
	Mechanical	Aerospace Ground Equipment Mechanic
	Administrative	Personnel Specialist
	Administrative	Information Systems Radio Operator
	General	Air Traffic Control Operator
	General	Aircrew Life Support Specialist
	Electrical	Precision Measurement Laboratory Equipment Specialist
	Electrical	Avionic Communications Specialist
Marine Corps	General	Infantry (Rifleman, Machinegunner, Mortarman, Assaultman, Unit Leader)
	Mechanical	Automotive Mechanic
	Mechanical	Helicopter Mechanic

Instructions to Examinee

"During this test you must drive the tractor and semitrailer through the course. (Explain layout of course.) You do not have to perform PMCS [preventative maintenance checks]. You must perform each maneuver without assistance from me or the scorer assistant. You may adjust the seat and mirrors before we begin the test. Tell me when you are ready.

Operate Tractor and Semitrailer Task Score Sheet

	Steps	GO	NO-GO
1.	Selected first gear position.		
2.	Maintained RPM (did not allow vehicle to lug, jerk, or stall).		
3.	Shifted gears without grinding. (Mark N/A for M915)		
4.	Drove without riding clutch. (Mark N/A for M915)		
5.	Kept both hands on the wheel (except when shifting.		
6.	Used hand over hand steering when turning.		
7.	Passed restricted roadway without striking roadway markers.		
8.	Followed serpentine roadway.		
9.	Passed serpentine without striking barriers.		

NOTE TO SCORER: Do not allow soldier to stop before entering the restricted roadway and serpentine. Soldier should maintain 10-15 MPH going through both obstacles.

Figure 1. Example of Hands-On Performance Test Score Sheet

D. Using Maps/Following Proper Routes

How effective is each soldier in securing proper maps as needed; becoming familiar with routes ahead of time when appropriate; using maps effectively; following prescribed routes; and arriving at commitments on time?

Sometimes fails to secure or use maps when needed; may not be able to read maps properly; is often late i n reaching designated location. Sometimes fail to plan route ahead of time; may take unplanned route.	Almost always secures maps if needed and uses most maps effectively; usually completes commitments on time. Almost always plans route ahead of time; generally becomes familiar with route before commitment; usually follows planned route.	Always obtains proper maps when needed; is able to use grid coordinates to reach even hard-to-find sites; always completes commitments on time. Always plans and becomes familiar with route before commitments; always follows planned route.
1 2	3 4 5	6 7

Figure 2. Example of Behaviorally Anchored Rating Scale

Attachment A
Lessons Learned from Military Performance Assessment
(From Wise, 1992)

1. A careful specification of the domain to be assessed is critical. In the military arena there was a long history of job analysis that provided a good basis for defining the measurement domain. The biggest debates were over whether measurement should reflect maximal performance (labelled proficiency by the NAS Committee) or typical performance and be broadened to include aspects of effectiveness such as teamwork or performance under adverse conditions. Results showing significant examinee-by-task interactions also suggest the need to assure appropriate coverage of the target domain.

In the educational arena, the relative success in defining the domain of mathematics for the state-by-state comparisons was surprising to many. Building a consensus for other subjects may be somewhat more troublesome. The debate about whether and how to measure typical performance rather than just maximal performance is also very relevant. There is an explicit desire to create measures that reflect and hence motivate effort -- an exam the student can and should study for.

2. An understanding of the uses to be made of the measures should precede their design. The JPM project experienced tension between those who wanted to use the scores primarily to validate selection tests and those who also wanted to set performance standards that might be generalized across jobs. The former group was content with maximizing the reliable variability in individual differences. All they wanted was a good norm-referenced test. The latter group, wanted a carefully constructed, domain-referenced measure. The former group wanted an interval level scale for their correlational analyses. The latter group might have been satisfied with an ordinal scale for comparing individuals to standards, but actually argued for an absolute scale in order to simplify generalizations across jobs.

Multiple uses also have been discussed for improved educational achievement measures. If our purpose is to evaluate and improve educational systems, norm-referenced scales should suffice. If we want to address questions like how much education is enough (or how much should we spend on education), for either individuals or for the system as a whole, criterion-referenced scales may be more appropriate.

3. Attention should be paid to the type of scale that is required. The JPM performance scales had multiple uses. For some uses, such as dividing performance into acceptable versus unacceptable, an ordinal scale would suffice. For other purposes, such as placing dollar values on the utility of different performance levels, a ratio scale would be required. Item response theory (IRT) was developed in the educational arena. It provides a basis for generalizing scores across different sets of items or different samples of examinees. Many believe that it provides an "interval" scale of achievement. This is only technically true, within the framework of the measurement model. A different measurement model with monotonic transformations of the theta scale (and the associated item characteristic curves) could fit the data equally well. IRT models may not work well with more complex behavioral samples involving greater interdependencies among the scorable units. Also, a scale that suggests an "infinite" difference between knowing nothing and knowing a little may not be appropriate. The percent-GO, or percent-correct metric used in the JPM project provides a reasonable scale for competency assessment and should be considered in more complex educational assessment as well.

4. Written tests are not everything. Knowing facts about a procedure and being able to execute it successfully are not exactly the same thing. The results varied considerably across and within jobs, but generally sufficient differences were found between written and hands-on performance tests to encourage educational researchers to press on in efforts to develop alternatives. So far, however, very little has been changed as a result of developing and using alternative measures.

5. Alternatives measures are expensive to develop and even more expensive to administer and score. This finding is not at all new to those who have been working on standardized writing assessments. Difficulties

in developing equivalent prompts, in hiring and training scorers and the great time and costs associated with reading and scoring large numbers of essays are well known. In a time of great competition for educational dollars, the potentially modest benefits associated with better measures must be carefully weighed against the costs.

6. More attention might be given to analogs of performance ratings. Tests are one-time events. A "portfolio" approach offers some promise for assessing typical rather than just maximal performance, but well-developed rating scales and well-designed rater training programs may offer a lower cost alternative. Grades might be viewed as a form of rating, but frequently the desire for "objectivity" has removed any component of expert judgment. What is lacking is the type of standardization across teachers and schools that might be gained through carefully constructed and anchored scales. Other principles, such as the use of multiple raters and of rater training programs might also be integrated.

7. Procedures for assessing the generalizability of performance measures are important. Results from the JPM effort indicated very significant examinee-by-task interactions. Military personnel researchers are still debating the extent to which performance measures can be generalized across jobs. This also is a key issue in educational assessment. More novel performance tasks may not generalize well outside a specific domain of knowledge.